J Electr Eng Technol.2016; 11(6): 1857-1862
http://dx.doi.org/10.5370/JEET.2016.11.6.1857

ISSN(Print) 1975-0102
ISSN(Online) 2093-7423

Depth Images-based Human Detection, Tracking and Activity Recognition Using Spatiotemporal Features and Modified HMM

Shaharyar Kamal*, Ahmad Jalal† and Daijin Kim**

Abstract – Human activity recognition using depth information is an emerging and challenging technology in computer vision due to its considerable attention by many practical applications such as smart home/office system, personal health care and 3D video games. This paper presents a novel framework of 3D human body detection, tracking and recognition from depth video sequences using spatiotemporal features and modified HMM. To detect human silhouette, raw depth data is examined to extract human silhouette by considering spatial continuity and constraints of human motion information. While, frame differentiation is used to track human movements. Features extraction mechanism consists of spatial depth shape features and temporal joints features are used to improve classification performance. Both of these features are fused together to recognize different activities using the modified hidden Markov model (M-HMM). The proposed approach is evaluated on two challenging depth video datasets. Moreover, our system has significant abilities to handle subject's body parts rotation and body parts missing which provide major contributions in human activity recognition.

Keywords: Activity recognition, Depth camera, Feature extraction

1. Introduction

Recognizing human activities from video has made greater attention by researchers and become fundamental topic in pattern recognition research areas, including human machine interaction. Usually, the analysis is carried out by efficient feature extraction, learning and classification in order to compute the input patterns for recognizing activities [1-4]. Despite of undertaking research efforts and achieving significant results in the past decade with respect to human tracking and recognizing activities, there are still remain some challenges due to self-occlusion of human body parts, hidden body parts and fast human movements in complex background scenes. In addition, several researchers faced other problems in the form of light sensitivity and motion ambiguities due to conventional cameras.

To access high quality images and overcome the above mentioned problems, depth cameras [5-7] started new era for a variety of image recognition tasks including human activity recognition (HAR). These cameras facilitate to behave insensitive to lighting conditions, offering spatial characteristics and reducing body-occlusion. To review depth-based HAR research, Xia and Aggarwal [8] described an algorithm to extract interest points from depth videos and depth cuboid similarity feature (DCSF) to describe the local 3D depth cuboid for activity recognition. Jalal *et al.* [9] developed a novel life logging invariant features approach as 1D features profile to train and recognize different activities based on depth images. In [10], Oreifej and Liu proposed a new descriptor for activity recognition using a histogram capturing the distribution of the surface normal orientation in the 4D space of time and spatial coordinates. Also, combined features [11, 12] are used to analyze, train and recognize different activities. However, these methods either relied on the skeleton data or depth silhouettes data which causes low recognition accuracy especially in case of missing joint information, un-clear human silhouettes and large distance subjects. Therefore, we elaborate some novel features along with modified HMM to overcome the above mentioned problems and improve activity classification and accuracy.

In this paper, we presented a robust method to detect, track and recognize activities using depth silhouettes and body joints information. Firstly, we extract human silhouette using noisy background subtraction and floor removal techniques. Secondly, these depth silhouettes are extracted as depth shape and body joints features. These features are further symbolized. Finally, we train/recognize two depth datasets using modified HMM.

The outline of this paper is as follows: Section 2 presents the system architecture and further details of the proposed method. In Section 3, we explain the experimental results. Conclusion of the paper is discussed in Section 4.

† Corresponding Author: Dept. of Computer Science and Engineering, POSTECH, Korea. (ahmjal@yahoo.com)
* Dept. of Electronics and Radio Engineering, Kyung Hee University, Korea. (shka@khu.ac.kr)
** Dept. of Computer Science and Engineering, POSTECH, Korea. (dkim@postech.ac.kr)

Received: September 14, 2015; Accepted: May 16, 2016

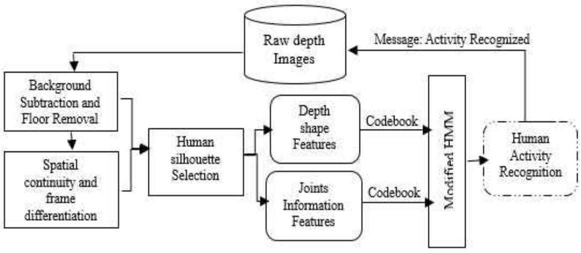

Raw depth Images
Message: Activity Recognized
Background Subtraction and Floor Removal
Spatial continuity and frame differentiation
Human silhouette Selection
Depth shape Features
Joints Information Features
Codebook
Codebook
Modified HMM
Human Activity Recognition

Fig. 1. System architecture of proposed activity recognition system

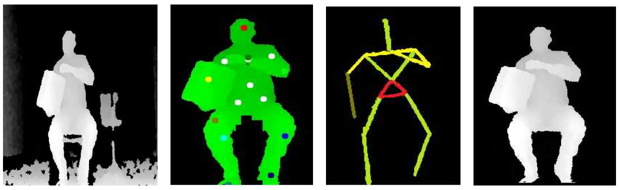


Fig. 2. Human silhouette selection by considering background subtraction and temporal differentiation

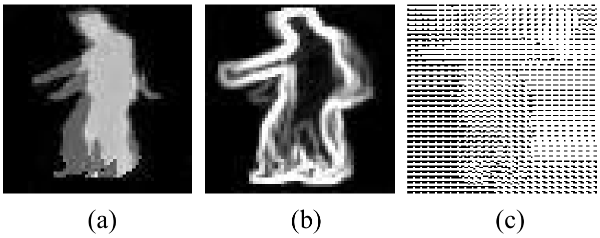


Fig. 3. (a) Depth image history; (b) temporal motion and (c) optical flow features applied over depth human silhouettes

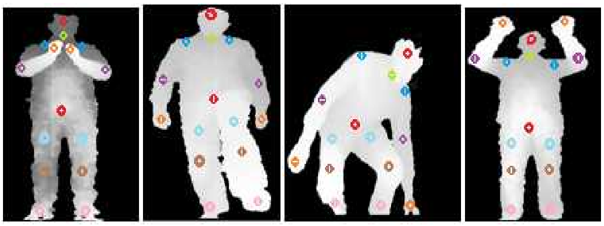


Fig. 4. Human body joints information

2. System Architecture and Methodology

The proposed activity recognition system is comprised on the tasks such as noisy background removal from raw depth data, human identification, feature extraction techniques, clustering using Linde, Buzo, and Gray (LBG)'s algorithm and training/recognition using modified HMM. Fig. 1 shows the overall flow of the proposed HAR system.

2.1 Depth image analysis

For the human silhouettes selection in depth sequential data, background subtraction routine is applied using least squares method. However, the floor is removed by ignoring the depth value y in a spaced grid [13]. While, all objects are localized via modified connected component labeling (see Fig. 2) and moving silhouettes are obtained by considering temporal depth differentiation as

$$F_{diff} = \sum_{i=1}^{n} \left| f_i^{(x,y,z)} - f_{i-1}^{(x,y,z)} \right| \rangle \, \tau \tag{1}$$

where depth silhouettes f deal with all three coordinates (x, y, z) with respect to successive frames i and i-1. τ is the specific threshold to evaluate the depth intensity values. To track moving human silhouettes, we considered the average of the disparity values in the detected parts and compare neighboring pixels surrounded by the detected moving parts [14] to make separate rectangular box for human identification.

2.2 Feature extraction using depth shape features

In depth shape features, firstly, depth image history features are obtained by considering the overall pixel intensity information of human body shape in sequential activities (see Fig. 3(a)). Secondly, temporal motion features are extracted by capturing the intra/inter motion variation among different body parts (see Fig. 3(b)). Thirdly, optical flow features are examined by considering the directional angular values among consecutive [15-17] frames (see Fig. 3(c)). All silhouettes are made more smoothen [18-20] for accurate results.

2.3 Feature extraction using joints information features

To analyze the joints information, we utilized our body part model that includes fifteen 3D body joints including head, neck, torso, shoulders, elbows, hands, hips, knees and feet as shown in Fig. 4. Hence, proposed real-time body parts tracking system [13] performed human pose estimation that utilize ridge body parts features to produce 15 different joints locations. During initialization, we used Y-shape to fixed different body constraints. Then, each body part provided their extremes values to recognize final pose in each activity frames.

In joint information features, firstly, angular joints features are measured by the difference of body parts angles among similar joints of consecutive frames as

$$\theta = \arccos\left(C_j^t - C_j^{t-1} \Big/ \left| C_j^t \right| \left| C_j^{t-1} \right| \right) \tag{2}$$

Secondly, body parts velocity features are captured by the movement of each joint in the direction of the normal vector of the plane from starting frame till ending frame [21, 22]. It is defined as

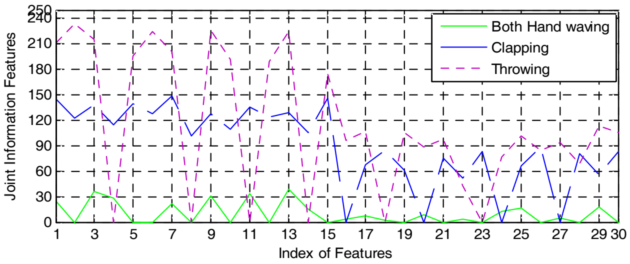


Fig. 5. 2D plot for joint information features

Table 1. Comparison of recognition accuracies using different codebook sizes of VQ and without VQ dataset.

Dataset	Codebook sizes		Without VQ
	128	256	
IM-DailyDepthActivity	**68.3**	63.9	66.1

$$V_j = \sum_{t=t_S}^{t_E} \left(v\left(j_t^{(x,y,z)}\right) - v\left(j_{t-1}^{(x,y,z)}\right) \right) \quad (3)$$

where, t_S and t_E are the starting and ending frames of sequential activity. However, the total feature dimensions obtained from the joints information features and body parts velocity features are 1x30 as shown in Fig. 5.

2.4 Sequence generation using vector quantization

Now, these depth and joint features vectors are further symbolized based on vector quantization (VQ) technique known as Linde, Buzo, and Gray (LBG)'s clustering algorithm [23]. We used the optimal codebook size of 128 after experimenting over different depth datasets. However, sequence of trained data get generated and maintained by buffer strategy [24, 25]. Table 1 shows comparison of different sizes of VQ and without VQ (i.e., actual dimensions) using our proposed dataset.

2.5 Modified Hidden Markov Model (M-HMM)

To train and recognize different activities, we applied the code vectors to the modified Hidden Markov models (M-HMM) having spatial and temporal variabilities. However, conventional HMMs include redundant information in the form of static body regions. Therefore, M-HMM is applied which mainly focused on active areas of human body characteristics such as moving feet and hips along with spatial/temporal properties of full-body shape of human silhouettes. During recognition phase of M-HMM, maximum likelihood value of specific sequential data is chosen to recognize distinct activity.

$$H_M = \arg\max_{l} \left\{ P(O \mid h_l) \right\} \quad (4)$$

where, $P(O \mid h_l)$ denoted the probability of likelihood of the h activity HMM among different number of activities [26-28]. Fig. 6 shows the internal structure of M-HMM.

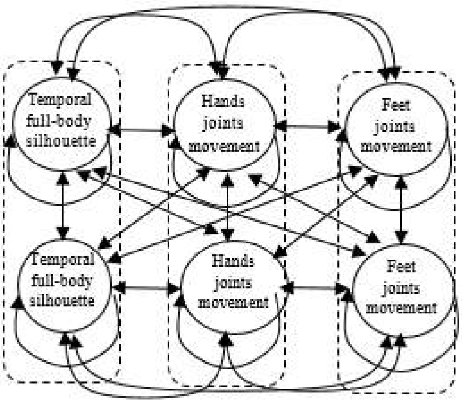
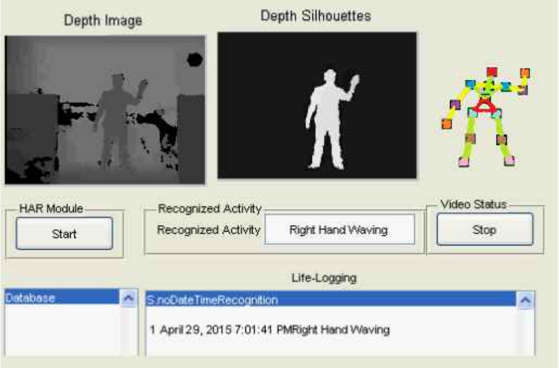


Fig. 6. Structural view of modified hidden Markov model



Recognized Activity: Right Hand Waving

S.noDateTimeRecognition

1 April 29, 2015 7:01:41 PMRight Hand Waving

Fig. 7. Proposed activity recognition system interface

3. Experimental Results and Analysis

In this section, we explain the interface of our proposed human activity recognition system and evaluate the proposed spatiotemporal features method over two challenging depth datasets.

3.1 Interface of proposed activity recognition system

During training/testing interface, we collected the raw depth data, extracted human silhouettes and manipulated the joints information. Then, random input activities were trained and recognized using modified HMM. Fig. 7 shows the overall concept of proposed HAR system.

3.2 Experiment I: MSRDailyActivity3D dataset

MSRDailyActivity3D dataset [29] is developed by Microsoft Research captured by a Kinect camera. This dataset includes 16 categories: drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa,

Table 2. Comparison of recognition accuracy using MSRDailyActivity3D dataset.

Methods	Recognition Accuracy
Joint position features [29]	68.0
Hybrid features [30]	73.8
Motion Variation features [31]	82.7
Robust features [38]	85.3
Actionlet ensemble [29]	85.7
Super normal vector [32]	86.2
Proposed method	91.3

Table 3. Comparison of recognition accuracy using IM-DailyDepthActivity dataset

Methods	Recognition Accuracy
Body Joints features [33]	28.4
Motion templates [34]	38.7
Eigenjoint feature [35]	40.3
Depth motion maps [4]	42.3
Super normal vector [32]	51.6
Depth silhouettes context [36]	57.6
Robust features [38]	58.7
Proposed method	68.3

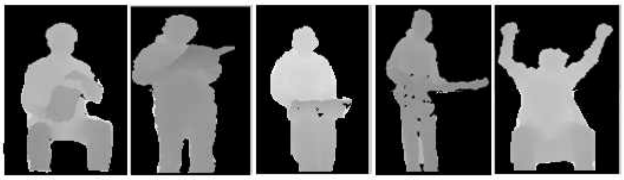


Fig. 8. Depth images of MSRDailyActivity3D dataset

walk, play guitar, stand up and sit down. The dataset consists of 320 video sequences which are performed by 10 subjects. Fig. 8 shows some depth images of MSRDaily Activity3D dataset.

To examine the recognition performance, we used cross subject validation process. In addition, we compared the experimental results of the proposed method with the algorithm defined as state of the art methods and the results are shown in Table 2.

3.3 Experiment II: IM-DailyDepthActivity dataset

To evaluate the online depth activity recognition scenarios, we build a new online activity dataset known as IM-DailyDepthActivity dataset [37] captured by depth camera. The dataset includes 15 different activities as: sit down, both hands waving, bending, stand up, eating, boxing, phone conversation, clapping, right hand waving, exercise, cleaning, kicking, throwing, take an object and reading an article performed by 15 subjects. The dataset is quite challenging due to similar postures of different activities. The total videos are 705, while 675 are training and 30 are testing. Fig. 9 shows some depth images of different activities used in IM-DailyDepthActivity dataset.

Table 3 summarizes the recognition accuracies of the proposed method along with state of the art methods. It is clearly analyzed that the proposed method achieved superior recognition accuracy than the existing methods.

Table 4. Comparison of computation time on SMMC-10 dataset.

Performance	Methods	
	Baak et. al [39]	Proposed
Computational time (fps)	60	289.9

Fig. 9. Depth images of IM-DailyDepthActivity dataset

Table 4 shows the computation time in terms of a speed of frames per second. It is clearly observed that the proposed method achieves the fast computation time and compared to well-known conventional method.

4. Conclusion

In this work, we proposed a novel methodology for activity recognition using spatiotemporal features and modified HMM via depth camera. The proposed system employed the depth body shape and joint information features, which are used to extract valuable feature vectors. These features are symbolized and trained/recognized using modified HMM. Experimental results showed that the proposed method using two depth activity datasets is superior in enhancing the recognition accuracy than other recommended approaches.

Acknowledgements

The research was supported by the Implementation of Technologies for Identification, Behavior, and Location of Human based on Sensor Network Fusion Program through the Ministry of Trade, Industry and Energy (Grant Number: 10041629). This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the “ICT Consilience Creative Program” (IITP-2015-R0346-15-1007) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

[1] X. Sun, H. Kashima and N. Ueda, “Large-scale personalization human activity recognition using online multitask learning,” *IEEE Trans. on knowl. and data engine.* vol. 25, no.11, pp.2551-2563, Nov.

2013.

[2] A. Jalal, Y. Kim, S. Kamal, A. Farooq and D. Kim, "Human daily activity recognition with joints plus body features representation using Kinect sensor," *in Proceedings of ICIEV Conference*, Fukuoka, Japan, pp.1-6, Jun. 2015.

[3] A. Jalal, N. Sharif, J. Kim and T. Kim, "Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart homes," *Indoor and Built Environment*, vol. 22, pp. 271-279, 2013.

[4] A. Jalal and I. Uddin, "Security architecture for third generation (3G) using GMHS cellular network," *in Proceedings of IEEE Conference on Emerging Technologies*, Islamabad, Pak, pp.74-79, Nov. 2007.

[5] A. Jalal, S. Lee, J. Kim and T. Kim, "Human activity recognition via the features of labeled depth body parts," *in Proceedings of ICOST Conference*, Artiminio, Italy, pp.246-249, June 2012.

[6] A. Jalal and S. Kamal, "Real-time life logging via a depth silhouette-based human activity recognition system for smart home services," *in Proceedings of AVSS Conference*, Seoul, Korea, pp.74-80, Aug. 2014.

[7] A. Jalal, S. Kamal and D. Kim, "A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments," *Sensors*, vol. 14, no. 7, pp. 11735-11759, 2014.

[8] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," *in Proceedings of CVPR Conference*, Portland, Oregon, pp.2834-2841, June 2013.

[9] A. Jalal, J. Kim and T. Kim, "Development of a life logging system via depth imaging-based human activity recognition for smart homes," *in Proceedings of Inter. Symposium on Sustainable healthy buildings*, Seoul, Korea, pp.91-95, 2012.

[10] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normal for activity recognition from depth sequences," *in Proceedings of CVPR Conference*, Portland, Oregon, pp.716-723, June 2013.

[11] A. Jalal, S. Kamal and D. Kim, "Shape and motion features approach for activity tracking and recognition from Kinect video camera," *in Proceedings of WAINA Conference*, Gwangju, Korea, pp.445-450, Mar. 2015.

[12] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal and D. Kim, "Robust human activity recognition from depth video using spatiotemporal multi-fused features," *Pattern Recognition*, 2016.

[13] A. Jalal and Y. Kim, "Dense depth maps-based human pose tracking and recognition in dynamic scenes using ridge data," *in Proceedings of AVSS Conference*, Seoul, Korea, pp.119-124, Aug. 2014.

[14] A. Farooq, A. Jalal and S. Kamal, "Dense RGB-D Map-Based Human Tracking and Activity Recognition using Skin Joints Features and Self-Organizing Map," *KSII Transactions on internet and information systems*, vol. 9, no. 5, pp. 1856-1869, 2015.

[15] A. Jalal, J. Kim and T. Kim, "Human activity recognition using the labeled depth body parts information of depth silhouettes," *in Proceedings of SHB symposium*, Korea, pp.1-8, Oct. 2012.

[16] A. Jalal and S. Kim, "Advanced performance achievement using multi-algorithmic approach of video transcoder for low bit rate wireless communication," *ICGST Journal of graphics, vision and image proc.*, vol. 5, no. 9, pp. 27-32, 2005.

[17] A. Jalal and M. Zeb, "Security and QoS Optimization for distributed real time environment," *in Proceedings of CIT Conference*, Aizu-Wakamatsu, Japan, pp. 369-374, Oct. 2007.

[18] A. Jalal and S. Kim, "The mechanism of edge detection using the block matching criteria for the motion estimation," *in Proceedings of HCI Conference*, Korea, pp. 484-489, Jan. 2005.

[19] A. Jalal and S. Kim, "A complexity removal in the floating point and rate control phenomenon," *in Proceedings of KMS Conference*, Korea, pp. 48-51, June 2005.

[20] A. Jalal, S. Kim and B. Yun, "Assembled algorithm in the real-time H.263 codec for advanced performance," *in Proceedings of Healthcom Conference*, Korea, pp. 295-298, June 2005.

[21] A. Jalal and A. Shahzad, "Multiple facial feature detection using vertex-modeling structure," *in Proceedings of ICL Conference*, Villach, Austria, pp.1-7, Sep. 2007.

[22] A. Jalal and M. Zeb, "Security enhancement for e-learning portal," *International Journal of Computer Science and Network Security*, vol. 8, no. 3, pp. 41-45, 2008.

[23] A. Jalal, M. Uddin, J. Kim and T. Kim, "Daily human activity recognition using depth silhouettes and R transformation for smart home," *in Proceedings of ICOST Conference*, Canada, pp.25-32, June 2011.

[24] A. Jalal and S. Kim, "Global security using human face understanding under vision ubiquitous architect-ture system," *World academy of science, engineering, and technology*, vol. 13, pp. 7-11, 2006.

[25] A. Jalal and Y. Rasheed, "Collaboration achievement along with performance maintenance in video streaming," *in Proceedings of ICL Conference*, Villach, Austria, pp.1-8, Sep. 2007.

[26] A. Jalal, M. Zia, J. Kim and T. Kim, "Recognition of human home activities via depth silhouettes and R transformation for smart homes," *Indoor and Built Environment*, vol. 21, pp. 184-190, 2012.

[27] A. Jalal, M. Zia and T. Kim, "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart

home," *IEEE Transaction on Consumer Electronics*, vol. 58, pp. 863-871, 2012.

[28] A. Jalal, Y. Kim and D. Kim, "Ridge body parts features for human pose estimation and recognition from RGB-D video data," *in Proceedings of ICCCNT Conference*, Hefei, China, pp.1-6, July 2014.

[29] J. Wang, Z. Liu, Y. Wu and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," *in Proceedings of CVPR Conference*, Providence, RI, pp.1290-1297, June 2012.

[30] S. Kamal and A. Jalal, "A hybrid feature extraction approach for human detection, tracking and activity recognition using depth sensors," *Arabian Journal for Science and Engineering*, pp. 1043-1051, 2016.

[31] A. Jalal, S. Kamal, A. Farooq and D. Kim, "A spatiotemporal motion variation features extraction approach for human tracking and pose-based action recognition," *in Proceedings of ICIEV Conference*, Fukuoka, Japan, pp.1-6, Jun. 2015.

[32] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," *in Proceedings of CVPR Conference*, Columbus, pp. 804-811, June 2014.

[33] A. Jalal, S. Kamal and D. Kim, "Depth map-based human activity tracking and recognition using body joints features and self-organized map," *in Proceedings of ICCCNT Conference*, Hefei, China, pp. 1-6, July 2014.

[34] M. Muller and T. Roder, "Motion templates for automatic classification and retrieval of motion capture data," *in Proceedings of ACM symposium on computer animation*, Austria, pp. 137-146, Sep. 2006.

[35] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-neartest-neighbor," *in Proceedings of CVPR Conference*, Providence, RI, pp. 14-19, June 2012.

[36] A. Jalal, S. Kamal and D. Kim, "Individual Detection-Tracking-Recognition using depth activity images," *in Proceedings of URAI Conference*, Goyang, Korea, pp. 450-455, Oct. 2015.

[37] A. Jalal, "IM-DailyDepthActivity dataset," imlab.postech.ac.kr/databases.htm, 2016, [Online; accessed February 5, 2016].

[38] A. Jalal, S. Kamal and D. Kim, "Depth Silhouettes Context: A new robust feature for human tracking and activity recognition based on embedded HMMs," *in Proceedings of URAI Conference*, Goyang, Korea, pp.294-299, Oct. 2015.

[39] A. Baak, M. Muller, G. Bharaj, H. Seidel and C. Theobalt, "A data-driven approach for real-time full body pose reconstruction from a depth camera," *in Proceedings of ICCV Conference*, Barcelona, Spain, pp. 1092-1099, Nov. 2011.

Shaharyar Kamal He is enrolled in Ph.D. degree in the Dept. of Radio & Electronics Eng. at Kyung Hee University, Korea. His research interest includes wireless communication, image processing.

Ahmad Jalal He received his Ph.D. degree in the Dept. of Biomedical Eng. at Kyung Hee University, Korea. His research interest includes human computer interaction, image processing, and computer vision.

Daijin Kim He received the Ph.D. degree in electrical and computer engineering from Syracuse University, Syracuse, NY. During 1992-1999, he was an Associate Professor in the Department of Computer Engineering at DongA University, Pusan, Korea. He is currently a Professor in the Department of Computer Science and Engineering at POSTECH, Pohang, Korea and Director of BK21+ POSTECH CSE Institute. His research interests include computer vision, human computer interaction, and intelligent systems.